Study and Application of Silence Model Adaptation for Use in Telephone Speech Recognition System

Jan NOVOTNÝ, Pavel SOVKA, Jan UHLÍŘ

Dept. of Circuit Theory, Czech Technical University, Technická 2, 166 27 Prague 6, Czech Republic

novotnj2@fel.cvut.cz, sovka@fel.cvut.cz, uhlir@fel.cvut.cz

Abstract. *This paper addresses the problem of the mismatch between a silence model and background noises which often occurs in a telephone speech recognition system (SRS) application. At first, the use of parallel model combination (PMC) methods is studied with the respect to this application. Secondly, the effective adaptation of a silence model to various background noises is confirmed. Finally, an original method combining log-add PMC with a noise power spectral density estimation based on minimum statistics is proposed. The performed tests prove the benefit of the suggested method to the speech recognition results that is caused by the stability of speech vector selection under the influence of various background noises. The advantages can be seen in no extra voice activity detector and in a relatively low computational load.*

Keywords

Robust speech recognition, silence model adaptation, parallel model combination.

1. Introduction

Study of robustness issues and their possible solutions is the important part of current speech recognition system development. Even more emphasis on this subject is needed when the recognition system is supposed to work in a real application like a telephone one [3]. This work follows up with [8] where a detailed study of telephone speech recognition system performance in noisy environment was introduced and it further develops the idea of silence model adaptation.

As the telephone nowadays can be used almost in any real environment, it results in a wide range of background noises occurring in the speech signal. The telephone communication habits of speaking persons (strength of voice, breathing, production of non-speech sounds, earpiece manipulation) are unfortunately unspecified too and this can be harmful for recognition results as well [7].

The non-speech signal production mentioned above leads into an inaccurate speech activity detection by SRS and consequently into the inadequate recognition results. Several methods of the solution of this problem were proposed [2, 3, 4, 5, 6, 7, 8]. A widely used solution is a voice activity detector (VAD) which discards non-speech signals prior to classification [5, 6]. Since the reliable VAD design and tuning is problematic [4, 6], an alternative approach employing silence model adaptation is proposed. An interesting solution developed to handle talker non-speech sounds using extended silence model was presented in [7]. It should be noted here that frequently used noise reduction methods (such as spectral subtraction) modify both speech and non-speech parts of the input signal. It was shown that just the modification of non-speech parts can bring a significant improvement to the speech recognition score [2]. Similarly, well-known model based techniques (such as parallel model combination [9, 10, 11, 12, 13]) compensate the silence model as well as speech models and thus the silence model compensation is seen as the important part of the methods. In the telephone communication, a speaking person talks close to the microphone and therefore the signal to noise ratio (SNR) of speech is relatively high. This is why just the problem of a silence model adaptation rather than speech models adaptation is studied in this paper and a practical solution is proposed and tested.

The paper is outlined as follows. The speech recognition system is briefly described in Section 2. Sections 3 and 4 describe and investigate several possibilities of a silence model adaptation with use of three different PMC techniques. A practical design and testing of silence model adaptation in conjunction with noise parameters estimation based on minimum statistics is presented in Section 5. Section 6 summarizes the results that were obtained.

2. Description of Speech Recognition System

An identical speech recognition system with [8] was utilized for evaluation of silence model adaptation. It means that the context dependent hidden Markov models (HMM) of phonemes trained on two thirds of the Czech database for the fixed telephone network (Speechdat(E)) [1] were used for the speech recognition system building. Mel-frequency cepstral coefficients analysis [16, 6] was applied, since it is a typical parameterization procedure linked with PMC methods.

The observation vector is composed of three streams. The first stream is represented by 12 static mel-cepstral coefficients and the 0'th cepstral coefficient. The second and

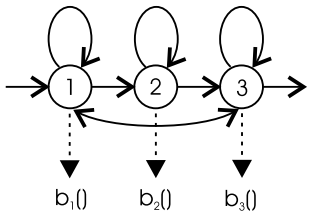

Fig. 1. The structure of silence model.

the third streams are composed of delta and acceleration coefficients respectively. The basic parameterization procedure setting is following. The input signal is windowed with a 32 ms Hamming window, the time shift between two following frames is set to 16 ms and the pre-emphasis coefficient is set to 0.97. The number of subsequent frames used for delta and acceleration coefficients computation is equal to 2 and cepstral liftering coefficient is set to 22.

The HMM training strategy is identical with [8], where more details about training, mel-frequency cepstral analysis optimization and the description of noise naturally present in the telephone communication can be found.

The HMM structure of silence model is shown in Fig. 1. The structure consists of three states. It slightly differs from the context dependent models of phonemes because it has forward-backward skip between the first and the last state. The mentioned skip allows the SRS to remain in silence model for longer time without unavoidable transition to the following word. The probability distributions of all states (both silence and context dependent phonemes) were divided into three streams and each stream was modeled as a three-component Gaussian mixture. The silence model was trained together with HMM of context dependent phonemes on two thirds of Speechdat(E) database, which means approximately 70 hours of records. This enabled various types of noise typical for telephone communication (impulse noises, breathing, transmission channel operation etc.) to be absorbed by the silence model.

3. Silence Model Adaptation Possibilities

Silence model adaptation possibilities are discussed in this section. The adaptation schemes generally utilize the assumption that ASR works best under the conditions in which it was trained (matched conditions) and thus they attempt to adapt once trained set of HMM to be adequate of hypothetic set of HMM trained in the matched conditions. In our case we try to adapt once trained silence model to be adequate to the current noise parameters. Well-known parallel model combination (PMC) algorithms [9, 10, 11, 12, 13] are seen to be well suited for this purpose. This is the reason why three variants of them were selected and reviewed with the respect of the application. A simplified general block scheme of PMC methods is depicted in Fig. 2.

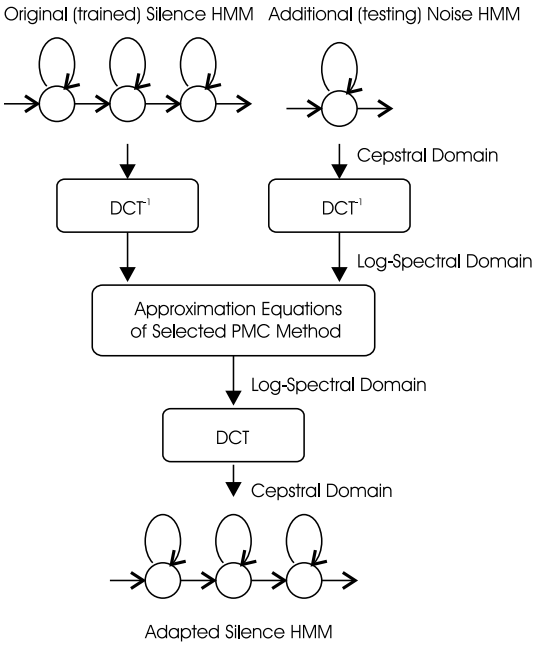

Fig. 2. Simplified general block scheme of PMC methods.

3.1 Silence Model Adaptation by Log-Normal PMC Method

Prior to the review of selected PMC methods, the utilized variables are summarized. Variables $\boldsymbol{\mu}^c$, $\boldsymbol{\Sigma}^c$ represent the mean vector and full covariance matrix, respectively, for models in the cepstral domain. All PMC compensation methods were used just for static parameters compensation; delta and acceleration were not compensated. In this case it means that the variables $\boldsymbol{\mu}^c$, $\boldsymbol{\Sigma}^c$ specify the probability distribution of the first stream (static parameters) and a selected mixture. Variables $\boldsymbol{\mu}^l$, $\boldsymbol{\Sigma}^l$ represent the mean vector and full covariance matrix again, but in this case for models in log spectral domain. Similarly $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ are the mean vector and full covariance matrix for models in linear spectral domain. Noise model parameters are specified by notation $\tilde{}$, the noise-compensated silence model parameters are represented by notation $\hat{}$ and the original silence model parameters (created in the training stage) have no additional notation. The subscripts $()_i$ and $()_{ij}$ are used for specification of a vector or a matrix component.

The first part of log-normal PMC [9] algorithm is a transformation of parameters from the cepstral domain to the log spectral domain. This transformation is performed by inverse DCT transform denoted by $\mathbf{C}^{-1}$ notation

$$\boldsymbol{\mu}^l = \mathbf{C}^{-1}\boldsymbol{\mu}^c, \quad \boldsymbol{\Sigma}^l = \mathbf{C}^{-1}\boldsymbol{\Sigma}^c\left(\mathbf{C}^{-1}\right)^T. \tag{1}$$

The next step is a conversion of log spectral parameters to linear spectral parameters. This is called an exponential transformation and described by the following equations

$$\mu_i = \exp\left(\mu_i^l + \Sigma_{ii}^l/2\right), \tag{2}$$

$$\Sigma_{ij} = \mu_i\mu_j\left[\exp\left(\Sigma_{ij}^l\right) - 1\right]. \tag{3}$$

The noise-compensated silence model parameters in the linear spectral domain are obtained by simple addition of the mean

vectors and full covariance matrices of both the original silence model and the noise model in linear spectral domains

$$\hat{\boldsymbol{\mu}} = \boldsymbol{\mu} + \tilde{\boldsymbol{\mu}}, \quad \hat{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma} + \tilde{\boldsymbol{\Sigma}}. \tag{4}$$

The noise-compensated silence model parameters need to be transformed back to log spectral domain. It is called logarithm transformation and expressed by the equations

$$\hat{\mu}_i^l = \log\left(\hat{\mu}_i\right) - \frac{1}{2}\log\left(\frac{\hat{\Sigma}_{ii}}{\hat{\mu}_i^2} + 1\right), \tag{5}$$

$$\hat{\Sigma}_{ij}^l = \log\left(\frac{\hat{\Sigma}_{ij}}{\hat{\mu}_i\hat{\mu}_j} + 1\right). \tag{6}$$

The last part of log-normal PMC compensation process is the conversion of models parameters from the log spectral domain to the cepstral domain. The process is made by DCT

$$\hat{\boldsymbol{\mu}}^c = \mathbf{C}\hat{\boldsymbol{\mu}}^l, \quad \hat{\boldsymbol{\Sigma}}^c = \mathbf{C}\hat{\boldsymbol{\Sigma}}^l\mathbf{C}^T. \tag{7}$$

3.2 Silence Model Adaptation by Log-Add PMC Method

Log-add PMC algorithm [11] is interpreted as a simplification of log-normal PMC, which assumes the variances are very small. In this case the covariance matrices are not compensated at all and this results in a significantly lower computational load. The assumption of very small or zero variances simplifies the mathematical notation of this method as well. The first part of log-add PMC algorithm is a transformation of mean vectors from the cepstral domain to log spectral domain

$$\boldsymbol{\mu}^l = \mathbf{C}^{-1}\boldsymbol{\mu}^c. \tag{8}$$

The compensation of static means in the log spectral domain can be easily described by the equation

$$\hat{\mu}_i^l = \log\left[\exp\left(\mu_i^l\right) + \exp\left(\tilde{\mu}_i^l\right)\right]. \tag{9}$$

The last part of log-add PMC compensation process is the conversion of mean vectors from the log spectral domain to the cepstral domain

$$\hat{\boldsymbol{\mu}}^c = \mathbf{C}\hat{\boldsymbol{\mu}}^l. \tag{10}$$

3.3 Silence Model Adaptation by DPMC Method

Data-driven PMC (DPMC) method [10] is based on the generation of synthetic observation vector sequences from original HMM parameters and from noise model parameters. These observation vector sequences in the cepstral domain are transformed to the sequences in the linear spectral domain by inverse procedure to parameterization. In this domain it is possible to add the given speech signal observation sequence (in our case given silence model observation sequence) and the additive noise observation sequence. Newly created observation sequences correspond to the hypothetic observation sequences in the selected noise and therefore they are used for new (adapted) estimation of HMM set parameters.

Mathematical description is similar to the previous ones. The notation is identical, but there are some new symbols. $\mathbf{O}(\tau)$ is an observation vector in time τ. Notation $\{\}_{\tau=1}^{T}$ means a sequence of length T. The first step of DPMC algorithm is the generation of observation vector sequences

$$\boldsymbol{\mu}^c, \boldsymbol{\Sigma}^c \Longrightarrow \{\mathbf{O}^c(\tau)\}_{\tau=1}^{T}. \tag{11}$$

The second step is the transformation of generated sequences from cepstral domain to linear spectral domain

$$\{\mathbf{O}(\tau)\}_{\tau=1}^{T} = \left\{\exp\left[\mathbf{C}^{-1}\mathbf{O}^c(\tau)\right]\right\}_{\tau=1}^{T}. \tag{12}$$

Newly created observation sequences are obtained by adding up the original observation sequences with the noise observation sequence

$$\left\{\hat{\mathbf{O}}(\tau)\right\}_{\tau=1}^{T} = \{\mathbf{O}(\tau)\}_{\tau=1}^{T} + \left\{\tilde{\mathbf{O}}(\tau)\right\}_{\tau=1}^{T}. \tag{13}$$

The new observation sequences are transformed back to the cepstral domain via standard parameterization procedure

$$\left\{\hat{\mathbf{O}}^c(\tau)\right\}_{\tau=1}^{T} = \left\{\mathbf{C}\left[\log\left(\hat{\mathbf{O}}(\tau)\right)\right]\right\}_{\tau=1}^{T}. \tag{14}$$

The mean vector and full covariance matrix in the cepstral domain for compensated models are computed by

$$\hat{\boldsymbol{\mu}}^c = \frac{1}{T}\sum_{\tau=1}^{T}\hat{\mathbf{O}}^c(\tau), \tag{15}$$

$$\hat{\boldsymbol{\Sigma}}^c = \frac{1}{T}\sum_{\tau=1}^{T}\left(\hat{\mathbf{O}}^c(\tau) - \hat{\boldsymbol{\mu}}^c\right)\left(\hat{\mathbf{O}}^c(\tau) - \hat{\boldsymbol{\mu}}^c\right)^T. \tag{16}$$

3.4 Discussion

It was stated in the previous text that the log-add PMC can be viewed as a simplification of log-normal PMC approach. The general disadvantage of this method is the incapability of covariance matrix compensation, but it does not play such an important role here because the covariance matrix of a background noise is difficult to be accurately estimated in the real application. The accuracy of the DPMC method is dependent on the length T of synthetic observation vector sequences. Thus in the real application a compromise between accuracy and computational load must be found. $T = 100$ was experimentally found as a good compromise in the following experiments.

4. Evaluation of Proposed Silence Adaptation Methods

At first, prior to the practical design and testing of silence model adaptation in the real environment, the best results of selected silence model adaptation methods were studied. The object was to test the SRS performance with and without the silence model adaptation methods under the presence of

	No silence model adaptation		Log-normal PMC silence model adapt.		Log-add PMC silence model adapt.		DPMC silence model adaptation		Log-normal PMC adapt.	
SNR	White	F1	White	F1	White	F1	White	F1	White	F1
[dB]	Corr/Acc	Corr/Acc	Corr/Acc	Corr/Acc	Corr/Acc	Corr/Acc	Corr/Acc	Corr/Acc	Acc	Acc
35	97.9/97.3	97.9/95.7	97.3/96.5	98.1/97.3	97.7/97.3	98.1/97.3	96.9/93.0	97.9/97.3	96.7	97.1
30	97.9/97.3	98.1/93.6	97.3/96.7	98.1/97.3	97.9/97.5	98.3/97.7	96.9/93.4	98.1/97.5	96.7	96.9
25	97.5/97.1	97.1/87.6	97.7/97.5	97.9/97.3	97.9/97.5	97.9/97.3	96.9/94.8	98.1/97.5	96.9	96.7
20	95.9/94.8	93.8/78.7	95.9/95.2	97.9/97.5	96.5/96.1	97.9/97.5	94.8/93.2	97.7/97.3	95.7	96.7
15	89.9/87.2	85.9/66.5	94.2/93.2	94.0/93.6	90.5/89.4	94.2/93.8	93.4/92.5	93.4/92.8	93.0	96.3
10	69.6/66.7	69.0/50.5	78.0/76.0	85.5/84.1	58.4/58.0	84.7/84.0	79.9/77.6	85.5/84.1	87.8	94.8
5	28.2/28.2	46.4/36.0	52.6/49.5	72.3/70.4	16.4/16.4	66.0/64.8	54.9/51.3	72.7/71.0	68.7	91.1
0	10.1/10.1	28.1/24.2	25.7/24.0	51.5/48.7	8.5/8.5	42.9/41.2	27.9/24.9	54.4/51.1	37.1	87.4
-5	9.9/9.9	16.4/15.1	11.0/11.0	31.9/29.4	9.5/9.5	24.6/23.4	11.4/11.4	33.7/30.8	13.2	81.6

	No silence model adaptation		Log-normal PMC silence model adapt.		Log-add PMC silence model adapt.		DPMC silence model adaptation		Log-normal PMC adapt.	
SNR	F2	F3	F2	F3	F2	F3	F2	F3	F2	F3
[dB]	Corr/Acc	Corr/Acc	Corr/Acc	Corr/Acc	Corr/Acc	Corr/Acc	Corr/Acc	Corr/Acc	Acc	Acc
35	97.7/96.5	97.5/95.0	97.5/96.1	97.9/96.7	97.7/96.7	97.9/97.3	97.7/96.1	97.7/97.5	96.3	96.1
30	97.7/95.7	96.1/89.8	97.5/96.3	97.5/97.5	97.5/96.3	97.9/97.9	97.5/96.5	97.9/97.9	96.3	97.1
25	95.4/91.7	93.0/84.9	95.7/94.6	96.3/96.3	95.9/95.0	95.9/95.9	96.1/95.2	96.1/96.1	95.5	96.5
20	87.4/80.8	84.3/72.9	93.0/92.1	93.6/93.4	92.7/91.9	93.8/93.8	92.8/92.1	94.0/94.0	94.2	96.3
15	72.5/66.5	68.7/59.2	85.9/84.5	91.7/91.7	86.3/85.5	90.7/90.7	86.1/85.5	91.1/91.1	94.0	95.9
10	50.9/46.6	49.7/43.7	72.7/71.4	87.4/87.4	72.9/71.4	84.5/84.3	72.3/70.6	87.2/87.2	90.1	95.5
5	26.5/25.1	30.2/27.9	60.4/57.3	78.7/78.7	59.0/56.3	75.2/74.8	58.8/56.5	82.6/82.6	90.7	95.9
0	15.3/15.3	14.3/13.5	46.2/43.9	63.4/62.7	41.8/38.7	55.1/54.5	43.9/42.0	75.4/75.2	89.9	95.2
-5	11.0/11.0	10.8/10.8	33.5/28.4	49.7/48.9	27.1/24.2	27.9/27.9	32.1/27.3	66.9/66.5	88.6	94.4

Tab. 1 Comparison of SRS results across the input SNR (the first column) when no model adaptation is used (the second wide column), the log-normal PMC silence model adaptation is performed (the third wide column), the log-add PMC silence model adaptation is applied (the fourth wide column), the DPMC silence model adaptation is performed (the fifth wide column) and the log-normal PMC adaptation of the whole model set is used (the last wide column). Subcolumns in the wide columns specify types of noise.

four different synthetically generated noises (white, F_1, F_2, F_3; description see below) with wide range of SNR and numerals dictation task [8].

4.1 Specification of a Testing Method and Evaluation of Results

The testing part of the database Speechdat(E) [1] was selected with the intention not to overlap with the training part of the database. Testing records (each from different person) contained ten Czech numerals zero to nine in random order. The length of the testing database was approximatelly 30 minutes. The pauses between the testing words have random duration which well simulates real applications. The extent of the database was set in order to be computationally feasible to test each SRS configuration in a quite extensive number of SNR values and noise types.

A synthetically generated stationary white noise and three types of stationary narrowband noises were used for testing of the robustness of SRS against the additional additive noises. The three types of stationary narrowband noises were generated by white noise filtering in purpose to affect the first (noise F_1, frequency band between 0.3 and 0.9 kHz), the second (noise F_2, frequency band between 1 and 2.5 kHz) or the third (noise F_3, frequency band between 2.5 and 3.4 kHz) formants. These noises are added to the testing speech records in order to achieve the specified SNR. The power of speech needed for SNR computation was obtained with the use of forced alignment (details can be found in [8]).

The results of the testing are the dependences of speech recognition performance on SNR of currently used testing noise. The speech recognition performance is evaluated by parameters acc (Percent Accuracy [%]) and $corr$ (Percent Correct [%]) [16]. The parameter $corr$ does not account for insertion errors, in this case for extra inserted commands. The insertion errors are often created during pauses and this is frequently related to inappropriate silence model setting. This is why the difference between $corr$ and acc parameters has been found useful for the SRS performance evaluation. If both mentioned parameters show similar results then it can be supposed that the extra inserted command error was suppressed and it consequently indicates the accurate silence model adaptation. Grammar scale factor and word insertion penalty values [16] were set to 5.0 and 0.0 respectively in all experiments as no significant improvement of results was achieved by attempt to optimize them in the tested task.

4.2 Results

The recognition results for five SRS configurations are presented in Tab. 1. The first configuration titled “No silence

model adaptation" is used as a reference. The results for this configuration are obtained with the use of the original silence model trained on Speechdat(E) database. It can be seen that the recognition system is able to operate well under the assumption of very high SNR ($SNR > 35$ dB) for noises not present during the training stage. If an unexpected type of noise occurs (like F_1, F_2, F_3) the SRS tends to decrease its recognition results rapidly even if the SNR parameter is still relatively high ($SNR > 20$ dB). Moreover the difference between $corr$ and acc is high that means an unstable performance of SRS during speech pauses. The second to the fourth SRS configurations in Tab. 1 presume selected PMC silence model adaptation. In this case the original silence model is adapted to the tested noise prior to the recognition. The results show appreciable improvement of SRS performance for all tested noises when $SNR > 15$ dB. The difference between $corr$ and acc is much lower than in the previous case and verifies the stability of SRS performance during speech pauses. The best results of the silence model adaptation almost in all ranges were obtained by log-normal PMC method. The DPMC method performed very well under the assumption of very low SNR. In comparison with the other methods the log-add PMC method yielded little worse results. On the other hand it is much computationally cheaper and it doesn't need the noise variance estimation. The last part of the table shows the results of SRS when the whole HMM set is adapted by the log-normal PMC. The main target of this part of the table is to compare the results of uncompensated SRS, the SRS with silence model adaptation and the totally compensated SRS for a wide range of SNR. It can be stated that the compensation of the whole model set is desirable if $SNR < 15$ dB and brings a benefit especially for narrowband noises when corrupted part of speech spectra can be substituted by the part uncorrupted by noise. The main disadvantage of the compensation of the whole model set is a relatively high computational load caused by the huge number of models to be adapted. A detailed analysis of computation complexity of PMC approaches can be found in [13].

5. Practical Design and Testing of Dynamic Silence Model Adaptation

In the previous sections the ideal configurations of the silence model adaptation by PMC methods were tested. It means that an accurate additional noise power spectral density (PSD) estimate was supposed. Unfortunately, in the real application like a telephone one, there is no direct access to the noise PSD. The suggested practical solution called dynamic silence model adaptation takes the advantage of noise PSD estimation based on minimum statistics [15]. Since the covariance matrix of the background noise is difficult to be accurately estimated, the log-add PMC method was applied for the silence model adaptation. The block scheme of the proposed dynamic silence model adaptation is presented in Fig. 3. Noise PSD is continuously estimated during the input signal parameterization and this estimate is used for the log-add PMC silence model adaptation. Because the spectral minima tracking procedure is used for the noise PSD estimation, the additional noise with lower power and more stationary character rather than speaker non-speech sounds (included in pre-trained silence model) is estimated. This is the very important fact for a correct silence model adaptation.

The dynamic silence model adaptation technique is mathematically described as follows. At first, the PSD of noise has to be estimated from noisy speech signal. For this purpose a minimum tracking procedure is utilized on smoothed noisy observation vectors in the power spectral domain. The recursive filter of the first order is used for the observation vectors smoothing

$$\overline{\left|\hat{\mathbf{O}}(\tau)\right|^2} = \alpha\overline{\left|\hat{\mathbf{O}}(\tau-1)\right|^2} + (1-\alpha)\left|\hat{\mathbf{O}}(\tau-1)\right|^2, \quad (17)$$

where α specifies the extent of smoothing. The PSD estimate of noise is obtained as the minimum within time interval T

$$\tilde{\mu}_i^2(\tau) = \beta \cdot \min\left\{\overline{\left|\hat{O}_i(\tau)\right|^2}\right\}_{\tau-T}^{\tau}, \quad (18)$$

where parameter β specifies the compensation of the final PSD estimate. Secondly, log-add PMC compensation equation is performed (similarly to eq. 9)

$$\hat{\mu}_i^l = \log\left[\exp\left(\mu_i^l\right) + \tilde{\mu}_i(\tau)\right], \quad (19)$$

where $\hat{\mu}_i^l$, μ_i^l and $\tilde{\mu}_i^2(\tau)$ are the compensated static means, the original (trained) static means and the noise PSD estimate respectively. The compensated features in log-energy domain are then transformed into the cepstral domain again via the DCT.

The suggested dynamic silence adaptation technique was tested with use of synthetic (F_1, F_3) and real background noises to simulate a real application performance. The SNR computation and evaluation of recognition performance is the same as in the previous section. Four groups of background noises were formed. The first and second group (labels "F_1, F_3" in Tab. 2) contain stationary synthetic narrowband noises F_1 and F_3. These noises are the same as the ones used in the previous section and allow us to compare the noise compensations with an ideal PSD knowledge and the dynamic silence model adaptation. The third group (label "office" in Tab. 2) contains several different noises produced by an air-conditioner, computer fans and a vacuum cleaner. The fourth group (label "car" in Tab. 2) contains various noises recorded in a car. The recognition results for two SRS configurations are presented in Tab. 2. The first one titled "No silence model adaptation" is used as a base for comparison. When no silence model adaptation is made then the results show an unstable SRS performance during speech pauses (the big difference between $corr$ and acc parameters, which is caused by a high word insertion error rate). If the proposed silence model adaptation was used (the column titled "Dynamic silence model adaptation") then the SRS performance and its stability during pauses were improved. The reliable SRS performance (recognition performance and accuracy are above 90%) is achieved for $SNR > 15$ dB in all four environmental groups ("F_1", "F_3", "office" and "car"). Presented results confirmed that the proposed dynamic silence adaptation is the robust and effective method especially for telephone applications and can be seen as the alternative solution to the VAD based feature vector selection.

	No silence model adaptation				Dynamic silence model adaptation			
SNR	F1	F3	Office	Car	F1	F3	Office	Car
[dB]	Corr/Acc	Corr/Acc	Corr/Acc	Corr/Acc	Corr/Acc	Corr/Acc	Corr/Acc	Corr/Acc
35	97.9/95.7	97.5/95.0	97.9/96.9	97.9/97.1	97.9/97.3	97.9/97.7	97.9/97.5	97.9/97.7
30	98.1/93.6	96.1/89.8	98.5/97.5	98.3/97.5	98.1/97.5	98.3/98.3	98.1/97.9	98.3/98.1
25	97.1/87.6	93.0/84.9	98.1/97.1	98.5/96.3	97.9/97.3	96.7/96.7	97.5/97.3	97.7/97.5
20	93.8/78.7	84.3/72.9	96.1/93.2	96.5/90.7	96.9/96.3	94.2/94.2	95.7/95.4	96.7/96.3
15	85.9/66.5	68.7/59.2	89.4/84.7	89.0/77.4	92.8/91.9	90.3/89.9	89.4/88.8	91.9/91.3
10	69.0/50.5	49.7/43.7	66.3/61.9	62.5/51.6	82.4/80.8	81.2/81.2	67.5/67.3	68.3/67.7
5	46.4/36.0	30.2/27.9	32.5/30.2	27.5/24.2	60.9/59.2	66.9/66.7	30.8/30.8	33.1/32.9
0	28.1/24.2	14.3/13.5	13.2/12.8	10.3/10.1	37.9/37.1	41.6/41.4	10.1/10.1	11.2/11.2

Tab. 2 Comparison of SRS results when no silence model adaptation and the dynamic silence model adaptation were applied in the real environment testing ($\alpha = 0.75, \beta = 2.5$).

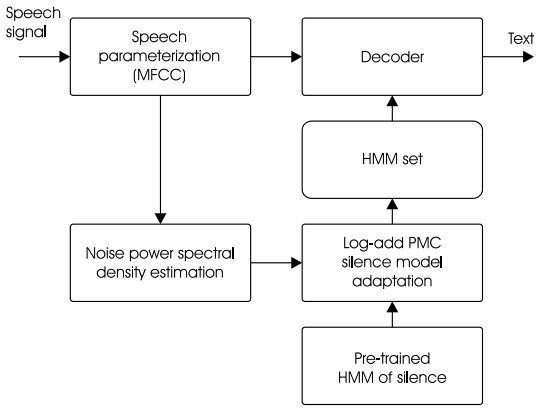

Fig. 3. Simplified block scheme of proposed dynamic silence model adaptation.

6. Conclusions

Various silence model adaptation possibilities were studied with the aim to make the SRS more robust. It was shown that the PMC methods are well suited for this purpose as they are able to adapt the originally trained silence model (which can well describe the talker's non-speech activity) to the unknown (and during training stage unseen) background noise. Well performed silence model adaptation improves stability and performance of SRS for the wide variety of background noises when $SNR > 15$dB.

A practical solution of dynamic silence model adaptation was designed. The original approach containing the noise estimation procedure with log-add PMC applied to noise model was suggested. Simulations confirmed a great benefit of this method to the SRS results and their stability in the supposed telephone applications. The main advantages of this solution are no extra voice activity detector and a relatively low computational load.

Acknowledgements

This work was supported by COST 278, EU Project, grant GAČR 102/02/0124: Voice Technologies for Support of Information Society and GAČR 102/03/H085.

References

[1] ČERNOCKÝ, J., POLLÁK, P., HANŽL, V. Czech Recordings and Annotations on CD's - Documentation on the Czech Database and Database Access. *Research Report*, Prague, CTU, 2000.

[2] HILGER, F., NEY, H. Noise level normalization and reference adaptation for robust speech recognition. In *Proc. ASR*, 2000, p. 64-68.

[3] RAMAN, V., RAMANUJAM, V. Robustness issues and solutions in speech recognition based telephony services. In *Proc. ICASSP*, 1997, p. 1523-1526.

[4] JUNQUA, J. C., MAK, B., REAVES, B. A robust algorithm for word boundary detection in the presence of noise. *IEEE Trans. on Speech and Audio Processing*, 1994, vol. 2, p. 406-412.

[5] VETH, J., MAUUARY, L., NOE, B., WET, F., SIENEL J., BOVES, L., JOUVET, D. Feature vector selection to improve ASR robustness in noisy conditions. In *Proc. Eurospeech*, 2001.

[6] *ETSI ES 202 050 V1.1.1*. European Telecommunication Standards Institute, 2002.

[7] ACCAINO, S., TSIPORKOVA, E., HAMME, H. Modeling of extra events for telephony. In *Voice operated telecom services: Do they have a bright future? workshop proceedings*, Ghent, Belgium, 2000, p. 75-78.

[8] NOVOTNÝ, J., SOVKA, P., UHLÍŘ, J. Analysis and Optimization of Telephone Speech Command Recognition System Performance in Noisy Environment. *Radioengineering*, 2004, vol. 13.

[9] GALES, M. J. F., YOUNG, S.J. HMM recognition in noise using parallel model combination. In *Proc. Eurospeech*, 1993, p. 837-840.

[10] GALES, M. J. F., YOUNG, S.J. A fast and flexible implementation of parallel model combination. In *Proc. ICASSP*, 1995, pp. 133-136.

[11] GALES, M. J. F. Predictive model-based compensation schemes for robust speech recognition. *Speech Communication*, 1998, vol. 25.

[12] HWANG, T. H., WANG, H. Ch. A fast algorithm for parallel model combination for noisy speech recognition. *Computer Speech and Language*, 2000, vol. 14, p. 81-100.

[13] HUNG, J., SHEN, J., LEE, L. New approaches for domain transformation and parameter combination for improved accuracy in parallel model combination (PMC) techniques. *IEEE Trans. on Speech and Audio Processing*, 2001, vol. 9, p. 842-855.

[14] MARTIN, R. Spectral subtraction based on minimum statistics. In *Proc. Eur. Signal Processing Conference*, 1994, p. 1182-1185.

[15] MARTIN, R. Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Trans. on Speech and Audio Processing*, 2001, vol. 9, p. 504-512.

[16] YOUNG, S. *The HTK Book (for HTK Version 3.1)*. Cambridge University Engineering Department, December 2001.